Reliable inference for complex models by discriminative composite likelihood estimation

Davide Ferrari*[1] and Chao Zheng[2]

[1,2]School of Mathematics and Statistics, University of Melbourne

December 15, 2015

Abstract

Composite likelihood estimation has an important role in the analysis of multivariate data for which the full likelihood function is intractable. An important issue in composite likelihood inference is the choice of the weights associated with lower-dimensional data sub-sets, since the presence of incompatible sub-models can deteriorate the accuracy of the resulting estimator. In this paper, we introduce a new approach for simultaneous parameter estimation by tilting, or re-weighting, each sub-likelihood component called discriminative composite likelihood estimation (D-McLE). The data-adaptive weights maximize the composite likelihood function, subject to moving a given distance from uniform weights; then, the resulting weights can be used to rank lower-dimensional likelihoods in terms of their influence in the composite likelihood function. Our analytical findings and numerical examples support the stability of the resulting estimator compared to estimators constructed using standard composition strategies based on uniform weights. The properties of the new method are illustrated through simulated data and real spatial data on multivariate precipitation extremes.

*Address: Richard Berry Building, University of Melbourne, Parkville, 3010, VIC, Australia; Phone: +61 383446411; E-mail: dferrari@unimelb.edu.au

Keywords: Composite likelihood estimation; Model selection; Exponential tilting; Stability, Robustness

1 Introduction

While likelihood-based inference is central to modern statistics, for many multivariate problems the full likelihood function is impossible to specify or its evaluation involves a prohibitive computational cost. These limitations have motivated the development of composite likelihood approaches, which avoid the full likelihood by compounding a set of low-dimensional likelihoods into a surrogate criterion function. Composite likelihood inference have proved useful in a number of fields, including geo-statistics, analysis of spatial extremes, statistical genetics, and longitudinal data analysis. See Varin et al. (2011) for a comprehensive survey of composite likelihood theory and applications. Larribe and Fearnhead (2011) review several applications in genetics.

Let X be a $d \times 1$ random vector and $f(x|\theta)$ be the assumed density model for X, indexed by the parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, $p \geq 1$. Suppose that the full likelihood function, $L(\theta|x) \propto f(x|\theta)$, is difficult to specify or compute, but we can specify low-dimensional distributions with one, two, or more variables. Specifically, let $\{Y_j, j = 1, \ldots, m\}$ be a set of marginal or conditional low-dimensional variables constructed from X with associated likelihoods $L_j(\theta|y_j) \propto f_j(y_j|\theta)$, where $f_j(\cdot|\theta)$, $\theta \in \Theta$ denotes the jth low-dimensional density model for Y_j. The low-dimensional variables $\{Y_j\}$ are user-defined and could be constructed by taking marginal models, like $X_1, \ldots, X_d$, pairs like (X_1, X_2), or conditional variables like $(X_1, X_2)|X_2$. The overall structure of such lower-dimensional models is sometimes referred as to composite likelihood design (Lindsay et al., 2011) and its choice is often driven by computational convenience. For example, if X follows a d-variate normal distribution $N_d(0, \Sigma)$, the full likelihood is hard to compute when d is large due to inversion of Σ, which involves $O(d^3)$ operations. In contrast, using sub-models for variable pairs $(X_k, X_{k'})$, $1 \leq k < k' \leq d$, can

reduce the computational burden since it involves simply inverting 2×2 partial covariance matrices.

Following Lindsay (1988), we define the composite likelihood function by

$$CL(\theta|w,x) = \prod_{j=1}^{m} f_j(y_j|\theta)^{w_j}, \tag{1}$$

where $\{w_j, j = 1, \ldots, m\}$ are non-negative weights, possibly depending on θ. A well-known issue in composite likelihood estimation is the selection of the weights, as their specification plays a crucial role in determining both efficiency and reliability of the resulting composite likelihood estimator (Lindsay, 1988; Joe and Lee, 2009; Cox and Reid, 2004; Varin et al., 2011; Xu and Reid, 2011). Despite the importance of the weights, many statistical and computational challenges still hinder their selection (Lindsay et al., 2011).

This paper is concerned with the aspect of stability of composite likelihood selection. Stability occurs when the maximizer of the overall composite likelihood function $L(\theta|w)$ is not overly affected by the existence of locally optimal parameters that work only for a relatively small portion of such sub-sets, say $Y_1, \ldots, Y_{m^*}$, $m^* < m/2$. The presence of such local optima arises from the incompatibility between the assumed full-likelihood model and the m^* lower dimensional models. For example, suppose that the true distribution of X is a d-variate normal distribution with zero mean vector, unit variance and correlations $2\rho_0$ for all variable pairs, while the true correlation is ρ_0 for some small fraction of the $d(d-1)/2$ pairs. If one mistakenly assumes that all correlations are equal to ρ_0, both maximum likelihood and pair-wise likelihood estimators with uniform weight, $w_j = 1/m$, $j = 1, \ldots, m$, are not consistent for ρ_0 in this situation. Other examples of incompatible models are given in Xu and Reid (2011). In applications, model compatibility is hard to detect, especially when m is large, so incompatible sub-models are often included in the composite likelihood function with detrimental effects on the accuracy of the global composite likelihood estimator.

Motivated by the above issues, we introduce the discriminative maximum composite

likelihood estimator (D-McLE), a new methodology for reliable likelihood composition and simultaneous parameter estimation. The new approach computes smooth weights by maximizing the composite likelihood function for a sample of observations subject to moving a given distance, say ξ, from uniform weights. The D-McLE is regarded as a generalization of the traditional McLE. If $\xi = 0$ the D-McLE is exactly the common composite likelihood estimator with uniform weights. When $\xi > 0$, incompatible sub-models are down-weighted, thus resulting in estimators for θ with bounded worst-case bias. Our analytical findings and simulations support the validity of the proposed method compared to classic composite likelihood estimators with uniform weights. The new framework is illustrated through estimation of max-stable models, which have proved useful for describing extreme environmental occurrences as hurricanes, floods and storms (Davison et al., 2012).

The proposed procure would be useful in two respects. First, the resulting weights would be a valuable diagnostic tool for composite likelihood selection. Small weights would signal suspicious models, which could be further examined leading to improved assumptions. Conversely, the method can be employed to identify influential data sub-sets for many types of composite likelihood estimators. Second, the estimates obtained by such method would be trustworthy at least for the bulk of the data sub-sets models (which are compatible with model assumptions). Clearly, assigning the same weight to all the models including the ones in strong disagreement with the majority of data would lead to biased global estimates, which can be an untrustworthy representations of the entire data-set.

The proposed method is a type of data tilting, a general technique which involves replacing uniform weights with more general weights. To our knowledge, this is the first work that introduces tilting for lower-dimensional data sub-sets within the composite likelihood framework. In robust statistics, tilting has been typically employed to robustify parametric estimating equations, or to obtain natural data order in terms of their influence Choi et al. (2000). Tilting has also been used to obtain measures of outlyingness and influence of data-subsets; e.g., see Hall and Presnell (1999); Critchley and Marriott (2004); Lazar

(2005); Camponovo and Otsu (2012). Genton and Hall (2014) use a tilting approach in the context of multivariate functional data to ranking influence of data subsets.

The rest the paper is organized as follows. In Section 2, we describe the new methodology for simultaneous likelihood selection/estimation; we give an efficient algorithm and introduce the compatibility plot, a new graphical tool to assess the adequacy of the sub-models. In Section 3, we study the properties of the new estimator and give its limit distribution. In Section 4, we provide simulated examples in finite samples confirming our theoretical findings. In Section 5, we illustrate the new procedure to the Tasmanian rainfall spatial data on multivariate precipitation extremes. In Section 6, we conclude and discuss possible extensions for $m \to \infty$. Proofs of technical results are deferred to a separate appendix.

2 Methodology

2.1 Composite likelihood selection

Given independent observations $X^{(1)}, \ldots, X^{(n)}$ from the true distribution $G(x)$, we construct the set of marginal or conditional low-dimensional observations $Y_j^{(1)}, \ldots, Y_j^{(n)}$, $j = 1, \ldots, m$, and define the weighted composite log-likelihood function

$$\ell_n(\theta|w) \equiv \sum_{j=1}^{m} w_j \ell_{nj}(\theta) \equiv \sum_{j=1}^{m} \frac{w_j}{n} \sum_{i=1}^{n} \log f_j(Y_j^{(i)}|\theta), \tag{2}$$

where $w = (w_1, \ldots, w_m)^T \in [0,1]^m$ are constants playing the role of importance weights. The weight w_j characterizes the impact of the jth sub-likelihood,

$$\ell_{nj}(\theta) \equiv n^{-1} \sum_{i=1}^{n} \log f_j(y_j|\theta),$$

on the overall composite likelihood function $\ell_n(\theta|w)$. We define incompatibility by assuming there is a global parameter, say $\theta_0 \in \Theta$, which suits most sub-models. Specifically, we assume

partial models $Y_j \sim f_j(y_j|\theta_j)$, where $\theta_j \neq \theta_0$ if $j \leq m^* < m/2$ (incompatible models) and $\theta_j = \theta_0$, if $m^* < j \leq m$ (compatible models).

Next, we introduce the D-McLE procedure for simultaneous discrimination of discordant models and parameter estimation. We propose to select the weight w_j to be small when, for a value of θ that is appropriate for the majority of the data sub-sets, the sub-likelihood function for the jth data sub-set, $\ell_{nj}(\theta)$, is small. To this end, w is regarded as a discrete distribution on m points and the discrepancy between w and the uniform distribution $w_{unif} = (1/m, \ldots, 1/m)$ is measured by the Kullback-Leibler divergence

$$D_{KL}(w, w_{unif}) = \sum_{j=1}^{m} w_j \log(m w_j), \tag{3}$$

where $0 \leq D_{KL}(w, w_{unif}) \leq \log m$. For a given parameter θ, data-dependent weights $w_n = w_n(\theta)$ are then chosen by solving the following program

$$\max_{w} \{\ell_n(\theta|w)\}, \quad \text{s.t.:} \quad D_{KL}(w, w_{unif}) = \xi, \ \sum_{j=1}^{m} w_j = 1. \tag{4}$$

Finally, the D-McLE, denoted by $\theta = \hat{\theta}_\xi$, is then defined as the maximizer of the composite log-likelihood function

$$\ell_n(\theta) \equiv \ell_n(\theta|w_n(\theta))$$

where $w_n(\theta) = (w_{n1}(\theta), \ldots, w_{nm}(\theta))^T$ is the vector of data-dependent weights. Equivalently, $\hat{\theta}_\xi$ can be obtained by computing the profiled estimator $\hat{\theta}(w)$ by maximizing $\ell_n(\theta|w)$ for a given weight and then solve (4) with $\theta = \hat{\theta}(w)$.

The composite likelihood estimator $\hat{\theta}_\xi$ entails moving away from uniform weights in the direction that emphasizes the contribution of the most useful data sub-sets. If $\xi > 0$, the relative importance of the sub-likelihoods that are incompatible with the data is diminished in the composite likelihood equation (2). The special case when $\xi = 0$ corresponds to the composite likelihood estimator with uniform weights $w = w_{unif}$. Thus, all the data sub-sets

are regarded as equally compatible. Other divergence measures may be considered in place of the Kullback-Leibler divergence (3), which could be useful in particular estimation setups, although these are not pursued in this paper. The Kullback-Leibler divergence, however, has the advantage that allows one or more zero weights, and gives automatically nonnegative wights without imposing additional constraints by some algorithm to ensure this property. For example, when m is very large it could be useful to modify $D_{KL}(w)$ to promote sparsity, i.e. select relatively a large number weights that are exactly zero.

2.2 Data-adaptive weights and parameter estimation

The program in (4) is solved by maximizing the Lagrangian function

$$h(w, \lambda_1, \lambda_2|\theta) = \sum_{j=1}^{m} w_j \ell_{nj}(\theta) + \lambda_1 \{D_{KL}(w, w_{unif}) - \xi\} + \lambda_2 \left(\sum_{j=1}^{m} w_j - 1 \right), \tag{5}$$

where λ_1 and λ_2 are Lagrange multipliers. It is easy to see that the solution to (5) has the form

$$w_{nj}(\theta) \equiv \alpha_2 \exp\{\alpha_1 \ell_{nj}(\theta)\}, \quad j = 1, \ldots, m, \tag{6}$$

where α_1 and α_2 depend on the Lagrange multipliers λ_1 and λ_2. From the two constraints in (4), $\alpha_1 \equiv \alpha_1(\theta)$ and $\alpha_2 \equiv \alpha_2(\theta)$ are obtained by solving

$$\xi = \alpha_1 \frac{\sum_{j=1}^{m} \exp\{\alpha_1 \ell_{nj}(\theta)\} \ell_{nj}(\theta)}{\sum_{j=1}^{m} \exp\{\alpha_1 \ell_{nj}(\theta)\}} - \log \sum_{j=1}^{m} \exp\{\alpha_1 \ell_{nj}(\theta)\} + \log m, \tag{7}$$

and $\alpha_2 = 1/\sum_{j=1}^{m} \exp\{\alpha_1 \ell_{nj}(\theta)\}$. The D-McLE $\hat{\theta}_\xi$ is then computed by maximizing $\ell_n(\theta) \equiv \ell_n(\theta|w_n(\theta))$.

Lemma 1 in the appendix shows that computing the D-McLE, $\hat{\theta}_\xi$, is equivalent to solving

the estimating equations

$$u_n(\theta) \equiv \nabla_\theta \ell_n(\theta) = \sum_{j=1}^{m} w_{nj}(\theta) u_{nj}(\theta) = 0, \tag{8}$$

where $u_{nj}(\theta) \equiv n^{-1} \sum_{i=1}^{n} u_j(Y_j^{(i)}, \theta)$ denotes the partial score function corresponding to the jth data subset. Thus, $u_n(\theta)$ is a weighted estimating equation involving the partial scores with weights depending on the data and θ. A small weight w_{nj} implies a modest contribution of the jth score, u_{nj}, to the overall composite likelihood equation. The constant ξ is regarded as a stability parameter which can be used to control for the relative impact of the incompatible lower-dimensional likelihoods. Particularly, if ξ is large incompatible models will receive a low weight, with a relatively small effect on the final parameter estimates. If $\xi = 0$, all the sub-models are treated equally in terms of the impact of corresponding sub-likelihoods in $u_n(\theta)$.

Equation (8) highlights the resemblance to estimating functions of classic robust M-estimators, whose main aim is to reduce the influence of outliers in the full likelihood function. Indeed, the approach followed here coincides with the robust estimation approach by Choi et al. (2000) in the particular case where: $n = 1$, $Y_1, \ldots, Y_m$ are independent and all sub-models f_j, $j = 1, \ldots, m$ are all identical to the full likelihood model, f. In general, however, the D-McLE is very different from Choi et al. (2000) and other similar robust methods. The main difference is that the weights $\{w_{nj}\}$ in (6) refer to variables $Y_1, \ldots, Y_m$, which are constructed by taking sub-sets of the original vector X and are possibly correlated; in robust M-estimation weights refer to independent observations on the original vector X. Thus, in our approach n observations corresponding to the jth data sub-set, namely $Y_j^{(i)}$, $i = 1, \ldots, n$, receive the same weight, w_{jn}. This reflects our need to control for the incompatibility of a portion of the sub-models, say $f_1, \ldots, f_{m^*}$, $m^* < m$, rather than reducing the effect of outlying observations with respect to the full model f.

2.3 Computing

The form of equation (8) suggests a simple algorithm to simultaneously compute weights and parameter estimates. At each step of the algorithm, we update weights based on previous parameter estimates and then compute a fresh parameter estimate using the new weights. Starting from an initial estimate, $\hat{\theta}^{(0)}$, we compute:

$$\hat{\theta}^{(t)} = \left\{ \theta : \ 0 = \sum_{j=1}^{m} \hat{w}_j(\hat{\theta}^{(t-1)}) u_{nj}(\theta) \right\}, \quad t \geq 1, \tag{9}$$

until convergence is reached. We consider a relative convergence criterion on the weights and stop iterating when $\|w_{nj}^{(t+1)} - w_{nj}^{(t)}\| / \|w_{nj}^{(t)}\| < \varepsilon$, where $\varepsilon > 0$ is some tolerance level. A practical advantageis that (9) is easy to implement when a basic composite likelihood estimator with fixed weights is already available.

In our numerical studies, the algorithm gave satisfactory performances. In all our examples convergence was reached in a few iterations and we noted that the computational cost does not increase much as m grows. This behavior makes the proposed algorithm well-suited to high-dimensional problems with a large number of sub-likelihoods and is shared by analogous iteratively re-weighted algorithms for M-estimation with well-established theory (e.g. see Arslan (2004)). Although we do not offer theoretical insight on the general theoretical behaviour of our algorithm, convergence results may be derived following an argument analogous to Basu and Lindsay (2004) in the context of iteratively reweighted procedures for minimum divergence estimators.

2.4 Compatibility profile plots (CPPs)

Let $\Pi(\xi) = (p_1, \ldots, p_m)$ be the arrangement of indices $\{1, \ldots, m\}$ implied by $w_{np_1}(\hat{\theta}_\xi) < \cdots < w_{np_m}(\hat{\theta}_\xi)$, where $w_{nj}(\hat{\theta}_\xi)$, $j = 1, \ldots, n$, are data-dependent weights computed by the algorithm in Section 2.3. The ordering $\Pi(\xi)$ induces an importance ranking for the submodels in terms of their compatibility with the true distribution generating the data. Based

on this ranking, a graphical tool is introduced, called a compatibility profile plot (CPP). The CPP traces the fitted weights, $w_{nj}(\hat{\theta}_\xi)$, $j = 1, \ldots, m$, as ξ moves away from zero and can be used to inspect the compatibility of individual sub-likelihoods. For instance, a sharp decrease of the first m^* weights from uniform weights $w_{unif} = (1/m, \ldots, 1/m)$, suggests that the first m^* sub-likelihoods are likely to be misspecified and a different model should be used for such components. The weights often exhibit diverging trajectories (see for example Figure 2) which may be used to determine a suitable value for the parameter ξ. For example, the plots help us pick a value of ξ corresponding to a sufficient degree of separation between compatible and incompatible models. Eventually, ξ reaches an equilibrium point where the trajectories are maximally separated. After equilibrium, $m-1$ weights cluster together again as they tend to 0, where a single weight converges to 1.

2.5 Selection of ξ

The stability parameter ξ tunes the extent to which we down-weight incompatible models, which is important to discuss. One approach is to select the tuning constant ξ closest to 0 (i.e., closest to uniform weights) such that the point estimates of the parameters of interest are sufficiently stable. If all the sub-likelihoods are compatible, $\xi = 0$ already gives stable estimates and moving away ξ is expected to have little impact on the estimates. In the presence of incompatible sub-likelihoods, values of ξ close to 0 tend give unstable estimates in terms of bias and variance, so we move ξ away from 0 until stability is reached. For example in Figure 2 (right), the correlation estimator $\hat{\rho}_\xi$ is far from the true correlation value of 0.5 when $\xi = 0$. As ξ moves away from zero, $\hat{\rho}_\xi$ changes rapidly until stability is reached when $\xi = 0.51$. The above discussion suggests a simple data-driven procedure to select ξ:

(1) Define an equally spaced grid $0 = \xi_0 < \xi_1 < \xi_2 < \cdots < \xi_r \leq \log m$.

(2) Starting from ξ_0 compute the correspondent point estimates, $\hat{\theta}_{\xi_i}$, $i = 0, \ldots, r$.

(3) Select the optimal value using the stopping rule $\hat{\xi} = \{\min\ \xi_i : \|\hat{\theta}_{\xi_i} - \hat{\theta}_{\xi_{i-1}}\| < \tau\}$, where $\tau > 0$ is some threshold value.

By definition, $\hat{\xi}$ is the value closest to 0 such that the variation of the point estimates is smaller than some acceptable threshold. Based on our simulations, a grid between $\xi_1 = 0$ and $\xi_r = -\log(1/2)$, with $\tau = 5\% \times \|\hat{\theta}_0\|$ typically works well and choices not too far from 0 already give considerable stability. If a very small portion of data sub-sets are incompatible, it may be useful to consider refinements of the grid near $\xi = 0$, such as, $\xi_i = (i/n)$, $i = 1, ..., r$.

3 Properties

3.1 Large sample behavior of $\hat{\theta}_\xi$ and standard errors

To emphasize reliability aspects, it is helpful to distinguish between the true process generating the data and the parametric model used for inference. Assume that X has distribution $G(x)$, while the true distribution for the sub-vector Y_j is denoted by $G_j(y_j)$. The density function of Y_j with respect to the dominating measure μ is denoted by $g_j(y_j)$. Let $\{F_j(y_j;\theta), \theta \in \Theta\}$ be a parametric family of distributions for Y_j and let $f_j(y_j|\theta)$ denote the corresponding densities with respect to μ. We assume that $f_j(y_j|\theta)$ is identifiable, i.e. for $\theta_1 \neq \theta_2$, $\mu[\{Y_j : f_j(Y_j|\theta_1) \neq f_j(Y_j|\theta_2)\}] > 0$, for all $j = 1, \cdots, m$.

The composite likelihood function (2) is correctly specified if there is a parameter $\theta_0 \in \Theta$ such that $f_j(y_j|\theta_0) = g_j(y_j)$ for all $1 \leq j \leq m$; when no such θ_0 exists then (2) is misspecified, meaning that it contains incompatible models. The optimal parameter, θ^*_ξ, is defined as the minimizer of the weighted composite Kullback-Leibler divergence

$$\theta^*_\xi = \underset{\theta\in\Theta}{\operatorname{argmin}}\ E_G\left\{\log \frac{g(X)}{\prod_{j=1}^m f_j(Y_j \in X|\theta)^{w_j}}\right\} = \underset{\theta\in\Theta}{\operatorname{argmin}}\ \sum_{j=1}^m w_j \ell_j(\theta), \tag{10}$$

where

$$\ell_j(\theta) \equiv -E_{G_j}\ell_{nj}(\theta) = -E_{G_j}\left\{\log f_j(Y_j|\theta)\right\}$$

is the cross-entropy between the true distribution G_j and the parametric sub-model $f_j(\cdot|\theta)$ and $w_j \equiv w_j(\theta) \equiv \alpha_2(\theta)\exp\{\alpha_1(\theta)\ell_j(\theta)\}$ $(j = 1, \ldots, m)$ here denote asymptotic weights computed as in Section 2.3 with $\ell_{nj}(\theta)$ replaced by $\ell_j(\theta)$. In the remainder of the paper, we assume that θ_ξ^* is the unique maximizer of (10).

Next, consistency and asymptotic normality of $\hat{\theta}_\xi$ are established. We note that standard M-estimation theory cannot be applied directly to equation (8) because the weights $\{w_{nj}(\theta), j = 1, \cdots, m\}$ in (4) depend on random averages; thus some additional care is needed to characterize the asymptotic behavior of $\hat{\theta}_\xi$.

Proposition 3.1 *Assume: (C1) θ_ξ^* is an interior point in Θ; (C2) $\sup_{\theta\in\Theta}|\ell_{nj}(\theta) - \ell_j(\theta)| \xrightarrow{p} 0$ as $n \to \infty$ $(j = 1, \ldots, m)$; and (C3) $\sup_{\theta\in\Theta}\ell_j(\theta) < \infty$ $(j = 1, \ldots, m)$. Then the maximum composite likelihood estimator $\hat{\theta}_\xi$ converges in probability to θ_ξ^* defined in (10).*

A direct consequence is Fisher-consistency of $\hat{\theta}_\xi$, i.e. under correct composite likelihood specification the optimal target value is $\theta_\xi^* = \theta_0$ for all ξ. This can be seen by taking the expectation of equation (8) with $\theta = \theta_0$:

$$E_G\left\{\sum_{j=1}^{m} w_{nj}(\theta)u_{nj}(\theta)\right\}\Bigg|_{\theta=\theta_0} = E_{w_n(\theta)}\left\{\sum_{j=1}^{m} w_{nj}(\theta)E_{G_j}u_{nj}(\theta)\,\Bigg|\, w_n(\theta)\right\}\Bigg|_{\theta=\theta_0} = 0, \qquad (11)$$

since $E_{G_j}u_{nj}(\theta_0) = 0$ if and only if $G_j(\cdot) = F_j(\cdot|\theta_0)$, for all $1 \le j \le m$. This means that the estimating equation (8) is solved by θ_0 regardless of the choice of ξ, since changing the latter affects only the weights $\{w_{nj}(\theta)\}$, but not the partial scores $\{u_{nj}(\theta)\}$. Section 3.2 discusses bias in the presence of incompatible sub-likelihoods.

Proposition 3.2 *Under conditions (C1) – (C3) in Proposition 3.1 and additional regularity conditions given in the Appendix, $\sqrt{n}(\hat{\theta}_\xi - \theta_\xi^*)$ converges in distribution to the p-variate*

normal $N_p(0, H_\xi^{-1} K_\xi {H_\xi}^{-1})$ *as* $n \to \infty$, *where* H_ξ *and* K_ξ *are the following* $p \times p$ *matrices*

$$H_\xi = \sum_{j=1}^{m} w_j^* \left[H_j(\theta_\xi^*) + \alpha_1^* E\{u_{nj}(\theta_\xi^*)\} E\{u_{nj}(\theta_\xi^*)\}^T \right], \quad K_\xi = Var \left\{ \sum_{j=1}^{m} w_j^* u_{nj}(\theta_\xi^*) \right\}, \quad (12)$$

$H_j(\theta) = E\{\nabla_\theta u_{nj}(\theta)\}$, $w_j^* = w_j(\theta_\xi^*)$ $(j = 1, \ldots, m)$, $\alpha_1^* = \alpha_1(\theta_\xi^*)$ *and expectations are with respect to* G.

The random weights, $\{w_{nj}(\theta)\}$, play a crucial role in determining the asymptotic behavior of $\hat{\theta}_\xi$. This feature is also found in model averaging, where parameter estimators obtained from different models, say $\hat{\mu}_S \in \mathcal{S}$, are combined into a global estimator $\hat{\mu} = \sum_{s \in \mathcal{S}} w_{nS} \hat{\mu}_S$, through random weights w_{nS} (Claeskens and Hjort, 2008, Chapter 7). The connection with model averaging is further highlighted by the normal location example in Section 4. Here the random weights converge in probability to constants; thus, the asymptotic variance takes the usual sandwich form and H_ξ, K_ξ can be consistently estimated analogously to Varin et al. (2011) with weights $w_{nj}(\hat{\theta}_\xi)$ $(j = 1, \ldots, m)$, computed as in Section 2.3. Resampling techniques such as jackknife and bootstrap may be also used.

3.2 Bias under incompatible models

In this section, we examine the first-order properties of our estimator in the presence of incompatible models. For clarity of exposition, in this section we consider the case where $\Theta \subseteq \mathbb{R}^1$, but analogous arguments can easily extended to the general case. To represent incompatibility, we assume heterogeneous parameters for the first m^* sub-models. Particularly, let $g_j(y_j) = f(y_j|\theta_j)$, $\theta_j \in \Theta$, $(1 < j \leq m)$, where θ_j, follows the drift model $\theta_j = \theta_\delta = \theta_0 + \delta$, if $j \leq m^*$, and $\theta_j = \theta_0$, if $m^* < j \leq m$. In addition, we assume that the Fisher information $H_j(\theta) = E_G\left[\partial^2 \log f_j(X;\theta)/\partial\theta^2\right]$, $j = 1, \ldots, m$, are bounded away from zero and infinity. A first-order Taylor expansion of u_{nj} and ℓ_{nj} in (8) about θ_j under suitable regularity conditions

gives

$$0 = \sum_{j=1}^{m} \exp\{\alpha_1(\hat{\theta}_\xi)\ell_{nj}(\hat{\theta}_\xi)\}u_{nj}(\hat{\theta}_\xi) \approx \sum_{j=1}^{m} \exp\{\alpha_1^*\ell_j(\theta_j)\} \left[\hat{\theta}_\xi - \theta_0 + \theta_0 - \theta_j\right] H_j(\theta_j).$$

Re-arranging the above expression leads to the following approximation for the bias

$$\hat{\theta}_\xi - \theta_0 \approx \frac{m^*\delta \exp\{\alpha_1^*\ell_1(\theta_\delta)\}H_1(\theta_\delta)}{m^* \exp\{\alpha_1^*\ell_1(\theta_\delta)\}H_1(\theta_\delta) + (m - m^*)\exp\{\alpha_1^*\ell_{n1}(\theta_0)\}H_1(\theta_0)} = \frac{\delta}{1 + C(\theta_0, \delta)},$$

where

$$C(\theta_0, \delta) = \frac{(m - m^*)H_1(\theta_0)}{m^*H_1(\theta_\delta)} \exp\{\alpha_1^*(\ell_1(\theta_0) - \ell_1(\theta_\delta))\} \geq \frac{c_1}{c_2}\left(\frac{m}{m^*} - 1\right) \exp\{\alpha_1^*(\ell_1(\theta_0) - \ell_1(\theta_\delta))\}.$$

Therefore, an approximate upper bound to the bias, $|\hat{\theta}_\xi - \theta_0|$, is

$$\text{Max-Bias}(\hat{\theta}_\xi|\delta) \equiv \frac{|\delta|}{1 + \frac{c_1}{c_2}\left(\frac{m}{m^*} - 1\right) \exp\left\{-\frac{\alpha_1^*\delta^2 H_1(\theta_0)}{2}\right\}}, \tag{13}$$

which is regarded as the worst-case bias under incompatible models. Clearly, when $\xi = 0$ (equivalently, $\alpha_1^* = 0$), the worst-case bias grows linearly in δ. When $\xi > 0$, Max-Bias$(\hat{\theta}_\xi|\delta)$ is bounded and the estimator $\hat{\theta}_\xi$ achieves bias control. Particularly, if $\delta = 0$ and all the models are compatible, then Max-Bias$(\hat{\theta}_\xi|\delta) = 0$. If δ is large, since the denominator in (13) dominates the numerator, the maximal bias decreases quickly to 0.

A second-order Taylor expansion of u_{nj} and ℓ_{nj} in (8) about θ_j (not shown here) can be used to derive an upper bound for the mean squared error. Analogously to (13), when $\xi = 0$ (equivalently, $\alpha_1^* = 0$), the worst-case mean squared error grows quadratically in δ. When $\xi > 0$, the maximal mean squared error is bounded, meaning that the estimator $\hat{\theta}_\xi$ achieves both bias and variance control. This theoretical understanding is confirmed by the numerical simulations in Section 4.

As an illustration, Figure 1 shows the maximal bias for the multivariate normal model

$X \sim N_m(\theta, I)$ with $\theta_j = \theta_0 + \delta$ where $\delta = 0$ if $j = 1, \ldots, m^*$, and $\theta_j = \theta_0$, if $m^* \leq j \leq m$. Clearly, the classic estimator with equal weights ($\xi = 0$) is very risky for this model, since the maximal bias can be potentially very large. This undesirable behavior can be easily avoided by setting $\xi > 0$. Thus, if the degree of incompatibility is strong ($|\delta| \to \infty$), the worst-case bias approaches zero. For intermediate cases where $|\delta| < \infty$ the bias remains bounded and can be controlled by tuning ξ.

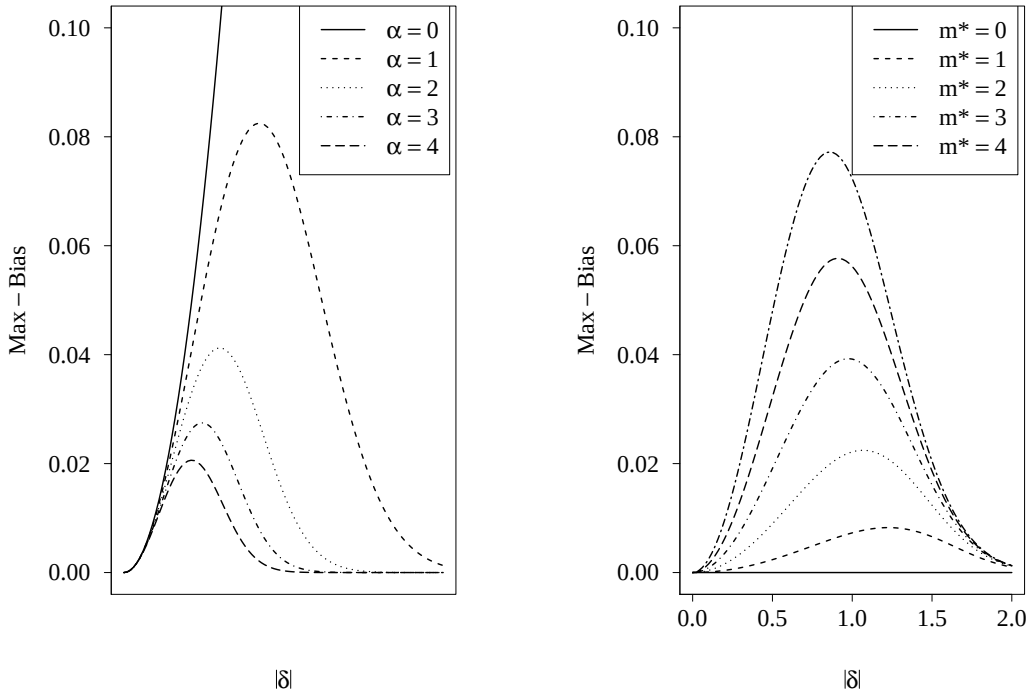

Figure 1: Worst-case bias for the multivariate normal location model $X \sim N_{10}(\theta, I)$ with $\theta_j = \theta_0 + \delta$ where $\delta = 0$, if $1 \leq j \leq m^*$, and $\theta_j = \theta_0$, if $m^* < j \leq 10$. Left: the curves correspond to different values of the constant α_1^* described in Sections 3.1 and 3.2 ($\alpha_1^* = 0, 1, 2, 3, 4$, and $m^* = 1$). Right: the curves correspond to increasing number of incompatible models, m^*, ranging from 0 (horizontal solid line) to 4 ($\alpha_1^* = 1$).

4 Examples

4.1 Example 1: Estimation of correlation

Suppose the random vector $(X_1, X_2, X_3, X_4, X_5)^T$ follows a multivariate normal distribution with zero mean vector, unit variances and covariances $Cov(X_1, X_k) = \rho_0/\sqrt{\varepsilon}$ if $2 \leq k \leq 5$, for some $\varepsilon \geq 1$, and $Cov(X_j, X_k) = \rho_0$ otherwise. If we model X as a multivariate normal with zero mean vector and all correlations equal, then the model is clearly misspecified and the maximum likelihood estimator is not consistent for ρ_0.

When constructing a composite likelihood function we only need pair-wise lower-dimensional likelihoods, since the marginal univariate sub-likelihoods do not contain information on ρ_0. Therefore the correlation estimator $\hat{\rho}_\xi$ is obtained as described in Section 2 by maximizing the pairwise likelihood

$$\ell_n(\rho|w) = \sum_{j>k} w_{jk}\ell_{njk}(\rho) \equiv \sum_{j>k} w_{jk}\left\{-\frac{n}{2}\log(1-\rho^2) - \frac{(SS_{jj}+SS_{kk})}{2(1-\rho^2)} + \frac{\rho SS_{jk}}{1-\rho^2}\right\}, \quad (14)$$

where $SS_{jj} = \sum_{i=1}^n (X_j^{(i)})^2$ and $SS_{jk} = \sum_{i=1}^n X_j^{(i)}X_k^{(i)}$. Note that (14) refers to combining bivariate normal models with zero mean and covariance given by 2×2 matrices with diagonal elements equal to 1 and off-diagonal elements equal to ρ. Therefore $\hat{\rho}_\xi$ will be consistent for ρ_0 only if $w_{n12} = \cdots w_{n15} = 0$.

In Table 4.1, we show the finite-sample bias and variance of the D-McLE for different values of ξ. As a comparison, we report results for the MLE and the usual McLE with uniform weights corresponds to the column with $\xi = 0$. When all the sub-likelihoods are compatible ($\varepsilon = 1$), not surprisingly the MLE has the best performance in terms of variance. For the D-McLE, however, both bias and variance do not increase much as long as ξ is not too far from 0. In the presence of incompatible sub-models ($\varepsilon = 3, 5$), the bias for the MLE and D-McLE with uniform weights ($\xi = 0$) is very large compared to the D-McLE with $\xi > 0$. For example, when $\varepsilon = 3$, the bias of the D-McLE is negligible when $\xi = 0.2$. In

addition to bias control of D-McLE, we note also that our procedure also achieves variance reduction when $\xi > 0$ and n is small. These results suggest that by setting ξ slightly above zero (e.g., 0.1, 0.2, or 0.3) already gives substantial stability and reduce the mean squared error of the corresponding estimator, $\hat{\theta}_\xi$.

	MLE	D-McLE(ξ)										
ε		$\xi=$ 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						Bias$^2 \times 100$						
1	0.00	0.00	0.06	0.14	0.21	0.27	0.37	0.43	0.52	0.62	0.68	0.78
3	2.43	1.75	0.12	0.00	0.05	0.12	0.20	0.26	0.36	0.43	0.50	0.58
5	6.32	4.53	0.46	0.00	0.05	0.11	0.19	0.27	0.34	0.42	0.51	0.57
						Var$\times 100$						
1	0.10	0.12	0.12	0.12	0.13	0.14	0.13	0.13	0.14	0.14	0.14	0.15
3	0.18	0.20	0.14	0.13	0.13	0.14	0.13	0.14	0.14	0.15	0.16	0.16
5	0.18	0.22	0.15	0.13	0.13	0.14	0.14	0.14	0.15	0.15	0.16	0.17

Table 1: Bias and variance for pairwise likelihood estimation of the correlation model $N_5(0, \Sigma)$ with unit variances and $Cov(X_1, X_k) = \rho_0/\sqrt{\varepsilon}$ if $2 \leq k \leq 5$, and $Cov(X_j, X_k) = \rho_0$ otherwise, with $\rho_0 = 1/2$ and $\varepsilon = 1, 3, 5$ ($\varepsilon = 1$ corresponds to the correctly specified model). The columns refer to maximum likelihood estimator (MLE) and the discriminative composite likelihood estimator (D-McLE) with ξ ranging from 0 to 1 ($\xi = 0$ implies uniform weights). Results are based on 10^4 Monte Carlo samples of size $n = 50$.

Figure 2 illustrates the profile plot (left) and parameter estimates (right) for a sample of $n = 50$ observations. When $\xi = 0$, the estimator is unreliable with estimates between $\rho_0/\sqrt{\varepsilon} = 0.5/\sqrt{5} \approx 0.22$ and $\rho_0 = 0.5$. When ξ moves away from zero, the importance profile shows two distinct groups of sub-likelihoods, with the four overlapping paths at the bottom corresponding to misspecified sub-likelihoods. When $\xi = 0.51$, the estimator $\hat{\rho}_\xi$ exploits correctly the information from the compatible sub-likelihoods and gives estimates close to the true value $\rho_0 = 1/2$. Finally, as $\xi \to \log(10)$, a single partial likelihoods tends to dominate the others, but much of the information from the other useful data pairs is ignored. Therefore the composite estimate at $\xi = \log(10)$ is inferior to that at $\xi = 0.51$, in terms of accuracy.

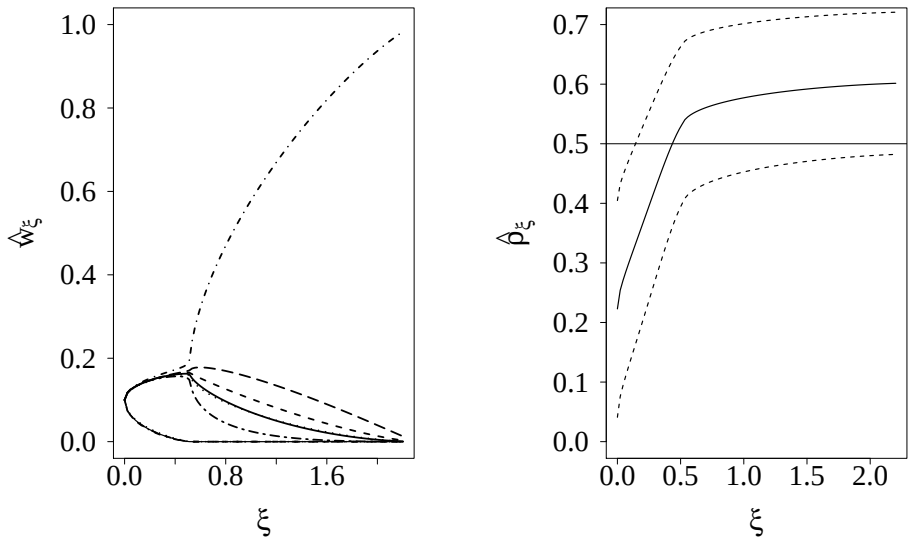

Figure 2: Estimation of the correlation model $N_5(0, \Sigma)$ with unit variances and $Cov(X_1, X_k) = \rho_0/\sqrt{5}$ $(2 \leq k \leq 5)$, and $Cov(X_j, X_k) = \rho_0$ $(j \neq k \neq 1)$, with true parameter $\rho_0 = 1/2$. Left: Importance profile paths for the partial likelihood components based on the estimated weights, $w_{n\xi}$. Right: estimated correlation coefficient (horizontal is the true value $\rho_0 = 0.5$). Illustration based on 50 observations.

4.2 Example 2: Location of heterogeneous normal variates

Let $(X_1, \ldots, X_m)$ be independent normal variables with common mean $E(X_j) = \mu_0$ $(1 \leq j \leq m)$ and heterogeneous variances $Var(X_j) = \sigma^2_{0,j}$ $(1 \leq j \leq m)$. This is the basic meta-analysis model where a weighted average of a series of study estimates, say $\{\overline{X}_j\}$, is combined to obtain a more precise estimate for μ_0. The inverse of the estimates' variance, $1/\sigma^2_j$, is the optimal study weight ensuring minimum variance of the combined estimate. All the parameter information is contained in the marginal models, so the following negative one-wise composite likelihood function is minimized:

$$-2\ell_n(\mu, \sigma_1, \ldots, \sigma_m | w) = \sum_{j=1}^{m} w_j \left\{ \log \sigma^2_j + \frac{1}{n} \sum_{i=1}^{n} \frac{(X_j^{(i)} - \mu)^2}{\sigma^2_j} \right\}. \tag{15}$$

and the profiled composite likelihood estimators are

$$\hat{\mu}(w) \equiv \sum_{j=1}^{m} w_j \overline{X}_j \equiv \sum_{j=1}^{m} \frac{w_j}{n} \sum_{i=1}^{n} X_j^{(i)}, \quad \hat{\sigma}_j^2(w) \equiv \frac{1}{n} \sum_{i=1}^{n} \{X_j^{(i)} - \hat{\mu}(w)\}^2, \quad j = 1, \ldots, m.$$

Replacing $\mu = \hat{\mu}(w)$ and $\sigma_j = \hat{\sigma}_j(w)$ in (15) gives $\sum_{j=1}^{m} w_j \log \hat{\sigma}_j^2(w)$, which is then minimized subject to the constraints $D_{KL}(w) = \xi$ and $\sum_{j=1}^{m} w_j = 1$.

The resulting location estimator, say $\hat{\mu}_\xi$, solves the fixed-point equation

$$\mu = \sum_{j=1}^{m} \hat{w}_j(\mu) \overline{X}_j = \frac{\sum_{j=1}^{m} \overline{X}_j \{\sum_{i=1}^{n} (X_j^{(i)} - \mu)^2\}^{-\hat{\alpha}_1}}{\sum_{j=1}^{m} \{\sum_{i=1}^{n} (X_j^{(i)} - \mu)^2\}^{-\hat{\alpha}_1}}, \tag{16}$$

where $\hat{\alpha}_1 > 0$ is computed as in (6) for a given $\xi \geq 0$, and the variance estimators are $\hat{\sigma}_{\xi,j}^2 = n^{-1} \sum_{i=1}^{n} (X_j^{(i)} - \hat{\mu}_\xi)^2$ $(j = 1, \ldots, m)$.

The degree of incompatibility of models is very strong, then the estimator $\hat{\mu}_\xi$ is nearly as good as the estimator obtained by ignoring the corresponding data sub-sets. If all the models are compatible, $\hat{\mu}_\xi$ still performs well in terms of accuracy. Particularly, if all the partial likelihoods are correctly specified, then $E(\hat{\mu}_\xi) = \mu_0$. If $\overline{X}_j$ $(1 \leq j \leq m^*)$ are far away from μ_0, then finding $\hat{\mu}_\xi$ is approximately equivalent to solving (16) with $w_{nj}(\mu) = 0$, if $j \leq m^*$.

Table 4.2 shows bias and variance for $\hat{\mu}_\xi$ under correctly specified and misspecified sub-likelihoods. The usual McLE with uniform weights corresponds to the column with $\xi = 0$. For comparison purposes, we also show the maximum likelihood estimator with weights $w_{mle,j} \propto 1/S_j^2$, where S_j^2 is the sample standard deviation for the jth variable. The results correspond to the location model with $\sigma_{0,j} = 1/j$ $(j = 1, \ldots, 10)$ and misspecification introduced by the location shift $\mu_j = \mu_0 + 1$, $j = 1, 2$. When all the sub-likelihoods are compatible ($m^* = 0$), the MLE has the best performance, but the D-McLE with $\xi = 0.1$ doing comparably well. In the presence of two incompatible sub-models ($m^* = 2$), the bias for the MLE and D-McLE with uniform weights ($\xi = 0$) is large compared to the D-McLE

with $\xi > 0$. The bias is quite small when $\xi = 0.3$. The variance of D-McLE for $0 < \xi \leq 0.3$ is also quite small compared to the McLE with uniform weights; interestingly in a few cases the variance is smaller than that of the MLE. This confirms the behavior observed in other numerical examples and in the derivations given in Section 3.2. Across a number of other simulation settings, we found that ξ slightly larger than zero gives estimators with negligible bias and relatively small mean squared errors.

		MLE	D-McLE(ξ)							
n	m^*		$\xi =$ 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
					Bias^2 $\times$ 1000					
10	0	0.00	0.01	0.00	0.01	0.07	0.14	0.19	0.21	0.22
	2	1.35	36.32	1.27	0.16	0.09	0.22	0.26	0.59	1.79
100	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	2.53	39.47	1.14	0.21	0.04	0.00	0.00	0.00	0.01
					Var$\times$1000					
10	0	3.51	5.03	4.08	6.69	9.55	11.01	11.89	12.49	12.90
	2	9.50	6.66	5.06	6.93	10.23	15.14	17.58	18.70	21.67
100	0	0.41	0.65	0.43	0.47	0.53	0.59	0.65	0.71	0.76
	2	0.53	0.73	0.48	0.51	0.53	0.55	0.57	0.59	0.61

Table 2: Bias and variance for location estimates of $X \sim N_{10}(\mu_0, \Sigma_0)$, where $\Sigma_0 = \text{diag}(1, 1/2, \ldots, 1/10)$, with and without incompatible models ($m^* = 0, 2$, respectively). The columns correspond to the maximum likelihood estimator (MLE) with weights proportional to $\{1/S_j^2\}$, where S_j^2 is the sample standard deviation for the jth variable, and the composite likelihood estimator with ξ between 0 and 0.7 (D-McLE). For $m^* = 2$, misspecification is introduced as $\mu_j = \mu_0 + 1$, $j = 1, 2$. Results based on 10^4 Monte Carlo samples of sizes $n = 10, 100$.

5 Multivariate models for spatial extremes: application to the Tasmanian rainfall data

Max-stable processes have emerged as a useful representation of extreme environmental occurrences such as hurricanes, floods and storms (Davison et al., 2012). However, their estimation poses significant challenges, since they lack of a general multivariate density expression. A well studied case is the Gaussian max-stable process defined as $Z(s) \equiv$

$\max_{i\geq 1}\{V_i f(U_i - s)\}$, where $\{V_i, U_i\}$ is a Poisson process on $(0,\infty] \times \mathbb{R}^2$, with intensity measure $\nu(ds) \times u^{-2}du$, and f is the bivariate normal distribution with zero mean and covariance Σ (Smith, 1990). The process Z has unit Frechét margins with distribution function $F(z) = \exp(-1/z)$, $z > 0$. Smith (1990) interprets Z as extreme environmental episodes, such as storms, where V, U, and f are the storm magnitude, center, and shape, respectively.

Next, we apply the D-McLE to estimate the extreme covariance parameter Σ in the context of the Tasmania rainfall data described below. For a finite set of spatially-referenced indexes, $s_1, \ldots, s_d \in \mathbb{R}^2$, the joint distribution of the random vector $Z(s_1), \ldots, Z(s_d)$ has no analytical representation for $d > 2$. Padoan et al. (2010) give a closed-form expression for the bivariate density and propose estimation based on the pairwise likelihood function. Given n observations on d locations, $z_1^{(i)}, \ldots, z_d^{(i)}$, $(i = 1, \ldots, n)$, the weighted pairwise likelihood function obtained by considering all $m(m-1)/2$ location pairs is

$$\ell_n(\Sigma|w) = \sum_{j=1}^{m-1}\sum_{k=j+1}^{m} w_{jk}\sum_{i=1}^{n}\log f_{Z_jZ_k}\left(z_j^{(i)}, z_k^{(i)}\middle|\,\Sigma\right),$$

where $f_{Z_jZ_k}$ is the bivariate density

$$f_{Z_jZ_k}(z_j, z_k|\Sigma) = \exp\left[\frac{\Phi\{g_1(h)\}}{z_j} - \frac{\Phi\{g_2(h)\}}{z_k}\right] \times \left\{\left[\frac{g_2(h)\varphi\{g_1(h)\}}{a(h)^2x_j^2z_k} - \frac{g_1(h)\varphi\{g_2(h)\}}{a(h)^2z_jx_k^2}\right]\right.$$
$$\left. + \left[\frac{\Phi\{g_1(h)\}}{x_j^2} + \frac{\varphi\{g_1(h)\}}{a(h)^2x_j^2} - \frac{\varphi\{g_2(h)\}}{a(h)^2z_jz_k}\right]\left[\frac{\Phi\{g_2(h)\}}{x_k^2} + \frac{\varphi\{g_2(h)\}}{a(h)^2x_k^2} - \frac{\varphi\{g_1(h)\}}{a(h)^2z_jz_k}\right]\right\}. \quad (17)$$

In the above expression, Φ and φ are the standard normal probability and density functions, respectively; $h = (s_j - s_k)$, $a(h) = (h^T\Sigma h)^{-1/2}$; $g_1(h) = a(h)/2 + \log(x_j/z_k)/a(h)$; and $g_2(h) = a(h) - g_1(h)$. For fixed h, the extremal dependence behaviour is determined by Σ, which is therefore the main interest for inference. Since the above model requires unit Frechét margins, the observed margins, y_j, are transformed in unit Frechét by the transformation $y_j = g_j(y_j) \equiv [1 + \zeta_j\{y_j - \mu_j\}/\gamma_j]_+$, where $u_+ = \max(0, u)$ and μ_j, γ_j and ζ_j are location, scale and shape parameters obtained from the empirical distribution.

We consider a data set of 20 yearly rainfall maxima recorded at 10 gauging stations from 1995 to 2014 in the Australian state of Tasmania corresponding to the following locations, also shown in Figure 3: Bushy Park, Ross, King Island, Eddistone Point, Geeveston, Strahan, Flinders Island, Marrawah, Rocky Point, Orford (source: `http://wwwc.bom.gov.au/tas/`). The max-stable Gaussian model is then fitted using a pair-wise likelihood function including all $m = \binom{10}{2} = 45$ pairs of locations. We compute estimates $\hat{\Sigma}_\xi$ for different choices of ξ ranging from 0 to $\log(45)$. Figure 3 (left) shows the a map of Tasmania with the 10 stations locations, and the edges denote fitted weights, $\hat{w}_{nj}$ corresponding to $\xi = 0.3$ (dashed lines represent weights smaller than the first quartile of fitted weights). Figure 3 (right) shows CPP plots for the weights. We note that pairwise likelihoods involving the King Island station (located at coordinates -39.88 , 143.88 on the map) exhibit a very weak degree of compatibility compared to locations in the southern and eastern part of the island. This suggest a different pattern for the precipitations for King Island in relation to the rest of the stations; thus pair-wise sub-models involving such a station should be further inspected and possibly revised.

Figure 4 shows estimated parameters $\hat{\sigma}_{11}$, $\hat{\sigma}_{12}$ and $\hat{\sigma}_{22}$ for values of ξ ranging from 0 to 1; the vertical bands represent 95% confidence intervals. For values of ξ larger than 0.3, the interval estimates appear quite stable. This can be seen by looking at the relative change in parameter estimate and also width of the confidence intervals. We can see that the estimated extremal correlation, $\hat{\rho}$, is notably affected by the measurements in a single station (King Island). As the sub-likelihoods involving that particular station receive increasingly low weights, the estimates change substantially. This behavior is consistent with that observed in our simulated data. To compare fitted models we also considered the composite likelihood information criterion for model selection discussed in Padoan et al. (2010) and defined by $CLIC(\xi) = -2\ell_n(\hat{\theta}_\xi) + \mathrm{tr}\{\hat{H}_\xi^{-1}(\hat{\theta}_\xi)\hat{K}_\xi(\hat{\theta}_\xi)\}$, where $\hat{J}_\xi$ and $\hat{K}_\xi$ are estimates of the matrices H_ξ and K_ξ defined in Section 3.1. We found that the $CLIC(\xi)$ decreases monotonically for ξ in $[0, 1]$ – particularly we have we have $CLIC(0) = 156.6$ and $CLIC(0.3) = 155.9$. This

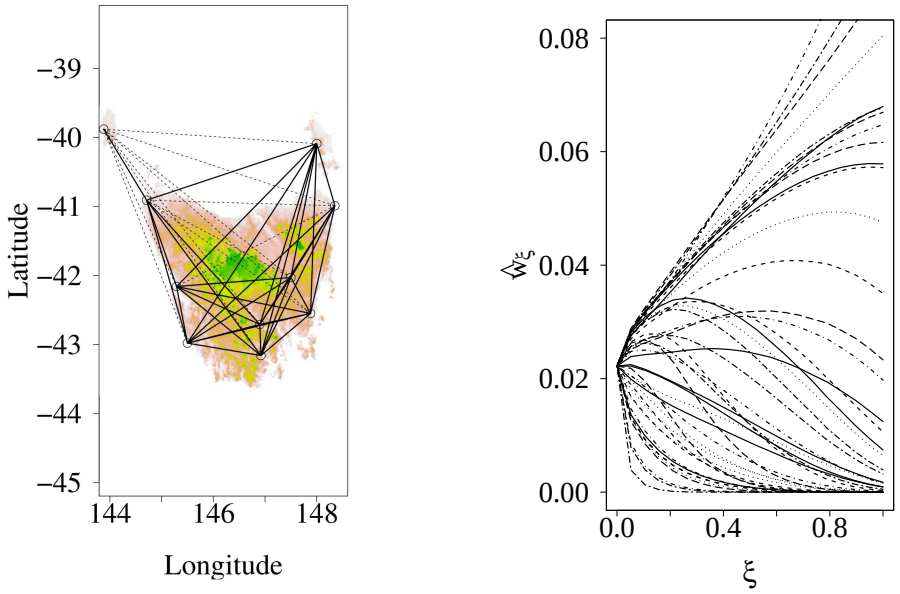

Figure 3: Left: Tasmania elevation map with location of the weather gauging stations. The dashed edges denote fitted weights w_{nj} (computed as described in Section 2.3) smaller than the first quartile for the weights (≈ 0.0065) when $\xi = 0.3$. Right: compatibility profile plots for ξ between 0 and 1.

suggests that $\xi > 0$ should be preferred to the usual composite likelihood estimator with uniform weights with $\xi = 0$.

6 Conclusion and final remarks

This work introduces the D-McLE, a new estimator obtained by maximizing the weighted composite likelihood function subject to a discrimination constraint, which entails moving away by a distance ξ from uniform weights. The D-McLE has appealing features from both theoretical and practical viewpoints. First, we found that the data-adaptive weights render the parameter estimates more stable in the presence of incompatible models compared to classic composite likelihood approaches with fixed weights. This is clearly seen from our asymptotic derivations and our numerical simulations confirm this behavior in finite samples. Second, the estimated weights, which are a by-product of our procedure, can be used to rank the compatibility of lower-dimensional likelihoods and are a useful diagnostic tool for model

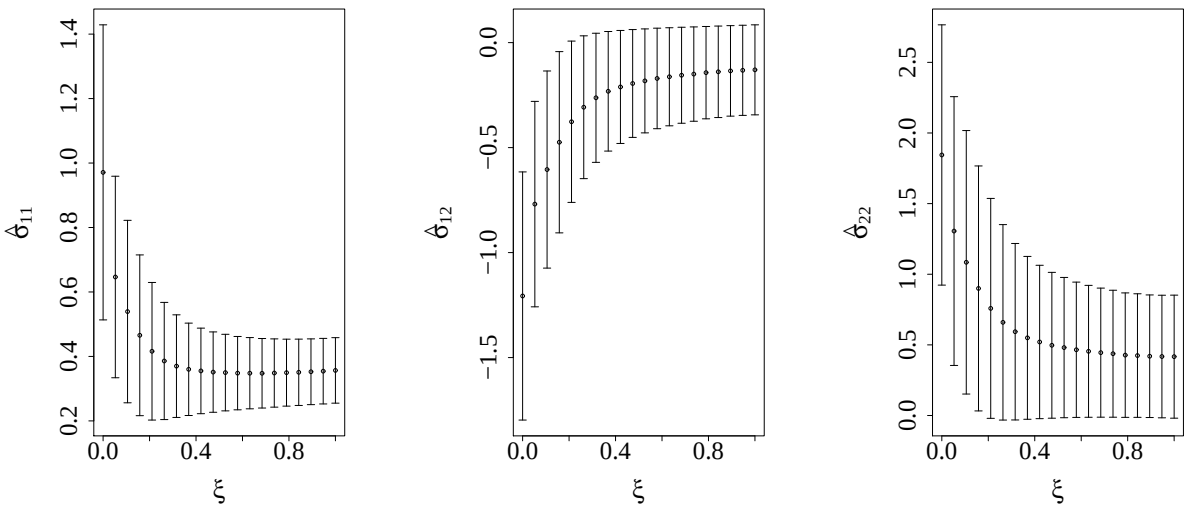

Figure 4: Estimation of the Gaussian max-stable model for the Tasmania rainfall data. Estimates of $\sigma_{11}, \sigma_{12}, \sigma_{22}$ for ξ ranging from 0 to 1. Vertical bars represent 95% confidence intervals based on standard errors from the asymptotic distribution of the D-McLE.

selection. For example, if the jth data sub-set receives an unusually small weight, it is likely that the corresponding model, $f_j(y_j|\theta)$, is incompatible. Targeted analyses on the anomalous data sub-sets can lead to improved model assumptions. Third, our approach leads naturally to the algorithm in Section 2.3, which we found to be quite fast and and easy to implement.

In recent years, high-dimensional estimation has become a core area of multivariate analysis. We believe that the D-McLE will be valuable as a remedy to common shortcomings of the classic McLE with fixed weights and MLE when the sample size, n, is relatively small compared to the complexity of the full model. Specifically, the constrained optimization problem (5) is a type of regularization approach where $\lambda_1 D_{KL}(w, w_{unif})$ can be regarded as complexity penalty which promotes sparsity and produces vectors w with many elements close to zero. Regularization approaches have proved useful for high-dimensional model selection (Fan and Lv, 2010). Similarly, in this context, we believe that the design of new sparsity-inducing penalty schemes for likelihood selection would be an interesting direction for further exploration and is high priority in our research agenda. Findings would be particularly valuable to spatial statistics and statistical genetics, where often the large number of

sub-likelihood components poses serious challenges to applicability of composite likelihood methods.

Up to date, not many papers have explored the large-m behavior of composite likelihood estimators from a theoretical perspective. Cox and Reid (2004) provide useful explanations on the asymptotic behavior of the pairwise composite likelihood estimator as $m \to \infty$ and n is fixed; particularly, they discuss how the presence of strongly correlated partial scores affects the usual convergence rate of McLE. Additional Monte Carlo experiments for the normal location model defined in Section 4.2 (not reported here) show that in finite samples the D-McLE can reduce considerably the mean squared error of the uniformly weighted McLE – even under fully compatible models. Such accuracy gains are relatively large when m increases. Developing theoretical insight on this phenomenon – and particularly on the interplay between the type of regularization constraint and the mean squared error of the resulting estimator as m increases – would represent another exciting future research direction.

Appendix

Lemma 1. If $D_{KL}(w_n(\theta), w_{unif}) = \xi$, $\xi \geq 0$, *then* $\nabla_\theta \ell_n(\theta) = \sum_{j=1}^m w_{nj}(\theta) u_{nj}(\theta)$. *Therefore,* $\nabla_\theta \ell_n(\theta) = 0$ *implies* $\nabla_\theta \alpha_1(\theta) = 0$ *with probability going to 1.*

Proof of Lemma 1. Let $\alpha_1'(\theta) = \nabla_\theta \alpha_1(\theta)$. Differentiating both sides of $D_{KL}(w_n(\theta), w_{unif}) = \xi$ gives

$$0 = \alpha_1'(\theta) \frac{\sum_{j=1}^m e^{\alpha_1(\theta)\ell_{nj}(\theta)} \ell_{nj}(\theta)}{\sum_{j=1}^m e^{\alpha_1(\theta)\ell_{nj}(\theta)}} + \alpha_1(\theta) \nabla_\theta \ell_n(\theta) - \frac{\sum_{j=1}^m e^{\alpha_1(\theta)\ell_{nj}(\theta)} \{\alpha_1'(\theta)\ell_{nj}(\theta) + \alpha_1(\theta) u_j(\theta)\}}{\sum_{j=1}^m e^{\alpha_1(\theta)\ell_{nj}(\theta)}},$$

where $\ell_n(\theta) = \sum_{j=1}^m w_{nj}(\theta)\ell_{nj}(\theta)$. This implies $\nabla_\theta \ell_n(\theta) = \sum_{j=1}^m w_{nj}(\theta) u_{nj}(\theta)$. A calculation

also shows

$$\nabla_\theta \ell_n(\theta) = \hat{\alpha}_1'(\theta) \sum_{j=1}^{m} w_{nj}(\theta)\{\ell_{nj}(\theta) - \ell_n(\theta)\}^2 + \sum_{j=1}^{m} w_{nj}(\theta) u_{nj}(\theta), \tag{18}$$

Since the first sum in (18) is strictly positive with probability one as $n \to \infty$ and the second sum equals zero by the Kullback-Leibler divergence constraint, we have that $\nabla_\theta \alpha_1(\theta) = 0$ with probability one as $n \to \infty$.

Proof of Proposition 1

The main goal is to show uniform convergence for the composite likelihood function $\ell_n(\theta)$. In particular,

$$\sup_{\theta \in \Theta} |\ell_n(\theta) - \ell(\theta)| \leq \sup_{\theta \in \Theta} \sum_{j=1}^{m} |w_{nj}(\theta)\ell_{nj}(\theta) - w_j(\theta)\ell_j(\theta)| \tag{19}$$

$$\leq \sum_{j=1}^{m} \sup_{\theta \in \Theta} |\ell_{nj}(\theta) - \ell_j(\theta)| + \sum_{j=1}^{m} \sup_{\theta \in \Theta} |\ell_j(\theta)| \; |w_{nj}(\theta) - w_j(\theta)| \tag{20}$$

The first term in (20) converges to zero in probability by Condition C2. By the continuous mapping theorem, also the second term converges to zero. Next, note that $\ell_n(\hat{\theta}_\xi) \geq \ell_n(\theta_\xi^*) = \ell(\theta_\xi^*) - o_p(1)$, where the last equality follows from the weak law of large numbers, since the latter implies $\ell_{nj}(\theta_\xi^*) \xrightarrow{p} \ell_j(\theta_\xi^*)$ $(1 \leq j \leq m)$, and the continuous mapping theorem. Hence

$$\ell(\theta_\xi^*) - \ell(\hat{\theta}_\xi) \leq \ell_n(\hat{\theta}_\xi) - \ell(\hat{\theta}_\xi) + o_p(1) \leq \sup_{\theta \in \Theta} |\ell_n(\theta) - \ell(\theta)| + o_p(1) \to 0, \tag{21}$$

by Condition C2. Since the optimal parameter θ_ξ^* value is unique, (21) implies $\hat{\theta}_\xi \xrightarrow{p} \theta_\xi^*$.

Regularity conditions and proof of Proposition 2

Let ∇ denote the differential operator with respect to the parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$, $u_n(\theta) = \sum_{j=1}^{m} w_{nj}(\theta) u_{nj}(\theta)$ denotes the weighted score p-vector, with partial scores $u_{nj}(\theta) =$

$n^{-1}\sum_{i=1}^{n}\nabla \log f_j(y_j^{(i)}|\theta)$, and H_ξ, K_ξ are $p\times p$ matrices defined in the asymptotic variance expression (12). Assume (C1)–(C3) given in Proposition 1 and the additional regularity conditions:

(C4) the sub-model $f_j(y_j|\theta)$ is three times differentiable in θ, $1\le j\le m$;

(C5) $\max_{1\le k\le m} E_G|u_{nk}(\theta)|^3$ is upper bounded by a constant;

(C6) the smallest eigenvalue of H_ξ is bounded away from zero;

(C7) the elements of the matrix K_ξ are upper bounded by a constant;

(C8) the expectation of second-order partial derivatives of $u_{nk}(\theta)$ with respect to G are upper bounded by a constant for all θ in a neighborhood of θ_ξ^*.

By Taylor's Theorem, there exists a random point $\tilde{\theta}$ between θ_ξ^* and $\hat{\theta}_\xi$ such that

$$0 = u_n(\hat{\theta}_\xi) = u_n(\theta_\xi^*) + \nabla u_n(\theta_\xi^*)(\hat{\theta}_\xi - \theta_\xi^*) + \frac{1}{2}(\hat{\theta}_\xi - \theta_\xi^*)^T\nabla^2 u_n(\tilde{\theta})(\hat{\theta}_\xi - \theta_\xi^*). \tag{22}$$

For the first term $u_n(\theta_\xi^*) = \sum_{j=1}^{m} w_{nj}(\theta_\xi^*)u_{nj}(\theta_\xi^*)$ in the above expansion, the central limit theorem implies that $\sqrt{n}\ u_{nj}(\theta_\xi^*)$ converges weakly to a p-variate normal distribution with mean $\mu_j^* = E_G u_{nj}(\theta_\xi^*)$ and $p\times p$ covariance matrix $V_j^* = -{H_j}^{-1}(\theta_\xi^*)$, for all $j = 1,\ldots,m$, where $H_j(\theta) = E_G\nabla u_{nj}(\theta)$. Since $\ell_{nj}(\theta_\xi^*) \xrightarrow{p} \ell_j(\theta_\xi^*)$ $(j = 1,\ldots,m)$, the continuous mapping theorem implies that $w_{nj}(\theta_\xi^*)$ converges in probability to constants $w_j^* = w_j(\theta_\xi^*)$ $(j = 1,\ldots,m)$. Therefore, by Slutsky's theorem we have convergence in distribution of $\sqrt{n}\ u_n(\theta_\xi^*)$ to the normal mixture

$$\sqrt{n}\ u_n(\theta_\xi^*) \xrightarrow{d} \sum_{j=1}^{m} w_j(\theta_\xi^*)N_p\{\mu_j^*, V_j^*\}.$$

such that $\sum_{j=1}^{m} w_j(\theta_\xi^*)\mu_j^* = 0$. For $\nabla u_n(\theta_\xi^*)$ in the second term of expansion (22), Lemma 1 gives

$$\begin{aligned}\nabla u_n(\theta_\xi^*) &= \sum_{j=1}^{m} w_{nj}(\theta_\xi^*)\left[\nabla u_{nj}(\theta_\xi^*) + \hat{\alpha}_1(\theta_\xi^*)u_{nj}(\theta_\xi^*)u_{nj}(\theta_\xi^*)^T + \{\nabla\hat{\alpha}_1(\theta_\xi^*)\}u_{nj}(\theta_\xi^*)^T\ell_{nj}(\theta_\xi^*)\right] \\ &\xrightarrow{p} \sum_{j=1}^{m} w_j^*\left\{H_j(\theta_\xi^*) + \alpha_1^*\mu_j^*\mu_j^{*T}\right\},\end{aligned}$$

where $\hat{\alpha}_1(\theta)$ is the solution of equation (6) and $\alpha_1^* = \alpha_1(\theta_\xi^*)$ denotes the solution of equation (6) with ℓ_{nj} replaced by ℓ_j and $\theta = \theta_\xi^*$. Convergence in probability follows from the continuous mapping theorem since $\ell_{nj}(\theta_\xi^*) \xrightarrow{p} \ell_j^*(\theta_\xi^*)$, $u_{nj}(\theta_\xi^*) \xrightarrow{p} \mu_j^*$, $\nabla u_{nj}(\theta_\xi^*) \xrightarrow{p} H_j(\theta_\xi^*)$. Finally, for the third term of the expansion (22) by assumption, there is a neighborhood B of θ_ξ^* and a constant κ for which each entry of the array $E_G\nabla^2 u_{nk}(\theta) < \kappa$ for all $\theta \in B$ and all $k = 1, \ldots, p$. Therefore, $\|\nabla^2 u_{nk}(\tilde{\theta})\|$ is bounded in probability by the law of large numbers. By Proposition 1, $\hat{\theta}_\xi \xrightarrow{p} \theta_\xi^*$ and the third term in the expansion (22) is of higher order than the second term, so the normality result follows by applying Slutsky's Lemma.

References

O. Arslan. Convergence behavior of an iterative reweighting algorithm to compute multivariate m-estimates for location and scatter. *Journal of Statistical Planning and Inference*, 118(1):115–128, 2004.

A. Basu and B. G. Lindsay. The iteratively reweighted estimating equation in minimum distance problems. *Computational statistics & data analysis*, 45(2):105–124, 2004.

L. Camponovo and T. Otsu. Breakdown point theory for implied probability bootstrap. *The Econometrics Journal*, 15(1):32–55, 2012.

E. Choi, P. Hall, and B. Presnell. Rendering parametric procedures more robust by empirically tilting the model. *Biometrika*, 87(2):453–465, 2000.

G. Claeskens and N. L. Hjort. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

F. Critchley and P. Marriott. Data-informed influence analysis. *Biometrika*, 91(1):125–140, 2004.

A. C. Davison, S. Padoan, M. Ribatet, et al. Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161–186, 2012.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

M. G. Genton and P. Hall. A tilting approach to ranking influence. *To appear in Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014. URL `http://stsda.kaust.edu.sa/Documents/2015.GH.JRSSB.pdf`.

P. Hall and B. Presnell. Biased bootstrap methods for reducing the effects of contamination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):661–680, 1999.

H. Joe and Y. Lee. On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, 100(4):670–685, 2009.

F. Larribe and P. Fearnhead. On composite likelihoods in statistical genetics. *Statistica Sinica*, 21(1):43, 2011.

N. A. Lazar. Assessing the effect of individual data points on inference from empirical likelihood. *Journal of Computational and Graphical Statistics*, 14(3):626–642, 2005.

B. G. Lindsay. Contemporary mathematics volume 80, 1988. volume 80, pages 221–239, 1988.

B. G. Lindsay, G. Y. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21(1):71–105, 2011.

S. A. Padoan, M. Ribatet, and S. A. Sisson. Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263–277, 2010.

R. L. Smith. Max-stable processes and spatial extremes. *Unpublished manuscript, University of Northern California*, 1990.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

X. Xu and N. Reid. On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, 141(9):3047–3054, 2011.